DOCUMENT RESUME

ED 286 926 TM 870 567

AUTHOR Adair, John G.; And Others
TITLE Hawthorne Control Procedures in Educational Experiments: A Reconsideration of Their Use and Effectiveness.
SPONS AGENCY Social Sciences and Humanities Research Council of Canada, Ottawa (Ontario).
PUB DATE Apr 87
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Control Groups; *Educational Research; *Effect Size; Experimental Groups; Matched Groups; Meta Analysis; *Research Design; *Research Problems; Sampling
IDENTIFIERS *Hawthorne Effect

ABSTRACT

A meta-analysis was conducted on 44 educational studies that used either a (labelled) Hawthorne control group, a manipulation of Hawthorne effects, or a group designed to control for the Hawthorne effect. The sample included published journal articles, ERIC documents or unpublished papers, and dissertations. The studies were coded on 20 variables, including the Hawthorne concern, Hawthorne control procedure, treatment controls, and sampling procedures. Each dependent measure was also coded on its outcome type. In most studies, the Hawthorne group was matched with the experimental group on a variable which reflected the Hawthorne concern: (1) an activity was performed to equate subjects with the experimental group's time, effort, interest, or novelty; (2) subjects were made aware that they were in an experiment; or (3) special attention was provided. Effect sizes were calculated for each treatment and for each Hawthorne group in comparison with the no-treatment control group, and for the treatment group in comparison with the Hawthorne group. Results in all but four cases were found to be nonsignificant, suggesting a database of generally poor quality. These preliminary results were consistent with previous reports of confusion surrounding artifact control. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED286926

Hawthorne Control Procedures in Educational Experiments: A Reconsideration of Their Use and Effectiveness

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

John G. Adair

Donald Sharpe

Cam-Loi Huynh

University of Manitoba

Winnipeg, Canada

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Adair

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April, 1987. This research was supported by a grant to the first author from the Social Sciences and Humanities Research Council of Canada.

TM 870 567

Years ago Desmond Cook (1962) raised the spectre of the Hawthorne effect as a particular kind of artifact that might contaminate experiments in education. His subsequent intensive research and unpublished conclusions (Cook, 1967), that there was a lack of understanding of the Hawthorne effect and that it might be less contaminating than at first thought, had far less impact than his earlier pronouncement. Moreover, Bracht and Glass (1968) had prominently listed the Hawthorne effect among contaminants influencing validity in educational experiments, thereby indelibly inscribing the artifact in many experimental methods texts.

A couple of years ago I reviewed this literature (Adair, 1984), and showed that the confusion surrounding Hawthorne and accompanying control procedures has persisted. Although struck by the diversity of control practices, I, as many others before, was searching for what might be called the "true" Hawthorne effect. My analysis led me to conclude that it was something akin to hypothesis awareness, a contaminant for which most control procedures were ill-suited.

However, one should not totally disregard the various attempts to control for Hawthorne that have been followed over the years. Each procedure represents a legitimate attempt to control for what was perceived to be of artifactual concern. Variations in the design strategies researchers select, in the concerns they seek to address, and in what they regard as an appropriate "placebo-like" activity for their Hawthorne Control group, may be a useful data source. By accumulating, coding and

categorizing these practices, the range and patterns of Hawthorne controls may be systematically described and meta-analytically examined to determine their effectiveness for each artifact source. We have undertaken this systematic examination of Hawthorne control practices and the present paper briefly reports on our progress and findings to date.

CLASSIFICATION OF RESEARCH PRACTICES

The project was limited to studies in education that employed either a "Hawthorne control group", a manipulation of Hawthorne effects, or a group by some other name for which the explicit rationale was the control of "Hawthorne effects". A total of 85 studies from the early 1960's to the present, were identified for review. Of the total sample, 29 were published journal articles, 31 were ERIC documents or unpublished papers, and 25 were dissertations. Although the database should be representative, it is far from exhaustive as many studies with comparable control practices were identified, yet not included because they failed to meet the criterion reference to Hawthorne.

Each study was coded on each of 20 separate dimensions, including such variables as the Hawthorne concern, the Hawthorne control procedure, the Hawthorne control activity, whether no treatment controls were also included, and sampling procedures. Each dependent measure was also coded on other dimensions, most importantly its outcome type, using a modification of the Smith, Glass, and Miller (1980) typology extended and adapted to educational research.

The range of Hawthorne control procedures we found and their

frequency of occurrence among the 85 studies are indicated in Table 1. In most studies the Hawthorne group was matched with the experimental group on one or more variables that reflected the artifact concern, either an activity to equate them with the experimental group in terms of time, effort, interest, or novelty, telling them that they were in an experiment, or giving them equivalent special attention. These three manipulations occured in various combinations, and occasionally along with manipulations to enhance the motivation of control subjects. Several non-matching procedures were also used, but it is clearly activity matching, awareness of participation and special attention that are the dominant Hawthorne manipulations.

What complicates an assessment of Hawthorne control procedures are the number of design-related dimensions on which studies varied. The frequency of each type of Hawthorne control group activity, and whether or not a no-treatment control group was included, are presented in combination with the Hawthorne procedures in Table 2. Along the left-hand column of the table, you can see that the tasks or activities given to Hawthorne controls varied widely, from something that is an irrelevant substitute for the experimental task (N=27), to activities that closely resemble the treatment in form or general nature (N=36), or that contain some substantive element (though not all) of the treatment procedure (N=14). This pattern applied both to studies with and without no-treatment control groups. Looking at the totals of activities in combination with the Hawthorne procedures reveals a substantial diversity of artifact control practices.

Moreover, any evaluation of Hawthorne procedures will be further complicated by other dimensions on which these practices systematically varied: for example, awareness Hawthorne procedures occured predominantly with non-random sampling, such as intact classes (78%) and with adult subjects, whereas approximately two-thirds of all attention Hawthorne controls occured with random sampling and with child subjects (68% and 65% respectively). In spite of these difficulties, the fact that we have been able to categorize and code each procedure on a number of relevant dimensions should provide us with a better handle with which to assess, in the next phase of our project, how the better procedures differ from the less effective ones.

META-ANALYSIS

The meta-analysis was undertaken on the 44 studies on which effect sizes could be calculated from among the 54 studies that included a no-treatment control group as a baseline. No studies were excluded on the basis of study quality.

Effect sizes were calculated for each treatment and for each Hawthorne group in comparison with the NTC, and for the treatment group in comparison with the Hawthorne group. An unbiased estimator of effect size, $\underline{d}$ (Hedges & Olkin, 1985), was used for all analyses. At each stage along the way, tests of the homogeneity of effect size were performed. In Table 3, the test statistics Q_{tc} and Q_{hc} (Hedges and Olkin, 1985, Chapter 6) were computed to assess the homogeneity of observations within each multiple-observation study under an outcome type. Under each outcome type, the homogeneity test statistics for observations

that were not significant within a multiple-observation study, as shown in Table 3, were pooled additively to constitute the value of Q_w, the within-study homogeneity test statistic, reported in Table 4. The between-study homogeneity test statistic, Qb, represents the difference between the value of an overall homogeneity test statistic and that of Q_w under each outcome type. The tests generally revealed homogeneity of observations within studies. In addition, studies were found to be homogeneous among outcome types, with the exception of OTs 11, 12, and 13.

To assess the effectiveness of Hawthorne procedures, the data were arranged into sets of effect sizes that included each appropriate comparison of treatment, Hawthorne, and NTC. For example, if a study contained two treatment groups and only one Hawthorne and one NTC group, there would be two comparison sets of treatment-Hawthorne-control comparisons, one for each treatment. There would be additional comparison sets for each dependent variable in studies containing more than one dependent variable. This process yielded 257 comparison sets, that constitute the basic observations computed from the sample of 44 studies. These data were arranged into comparison sets because of a fundamental assumption of our approach that the effectiveness of Hawthorne procedures could not be determined independent of their relationship to treatment and control procedures.

To assess the extent to which the observations under each outcome type yielded significant treatment or Hawthorne effects, each unpooled effect size was subjected to the null hypothesis

test (H_0: T-C = 0, H_0: H-C = 0, and H_0: T-H = 0). These significance tests could have yielded only one of eight possible patterns of significance-nonsignificance for each treatment-Hawthorne-control group set of comparisons. These are illustrated, described, and labeled in Figure 1, according to their fit to the conception of the Hawthorne effect and its presumed relationship to treatment and control groups. The ideal types are based on the view from the literature that an appropriately conceived Hawthorne group should manifest some greater effect than control subjects due to their special treatment, yet not so much of an effect that subjects perform comparable to those exposed to an experimental treatment. A perfectly positioned Hawthorne group would be one roughly equidistant from the treatment and NTC groups. The Strong Case is also regarded as an appropriate Hawthorne control. In this case the significant treatment effect is ordinarily ascribed to the Hawthorne effect. In the Weak Case the Hawthorne treatment is so inappropriately weak that these subjects perform essentially the same as NTC subjects. The Nonsignificant Case represents the study with totally nonsignificant results and hence nothing to interpret. The remaining cases also yield uninterpretable Hawthorne results since it is difficult to know what significant Hawthorne results mean when the presumably matching experimental treatments have yielded nonsignificant effects.

Frequency counts of the occurrence of each case of treatment-Hawthorne-control comparison are indicated in Table 5. The frequency of occurrence of significant effects were

distributed predominantly in nonsignificant cases (72%), 17% in the combined ideal and strong Hawthorne cases, with the least, only 11%, found among the combined weak and rare cases. Although, they are not reported in any table, all but 4 cases were found to be nonsignificant when these same tests were applied to the homogeneous weighted effect sizes for each study within each outcome type, and no significant cases were found among effect sizes pooled across studies within each outcome type. Clearly the large quantity of poor quality studies, when pooled together, overwhelmed any significant results that might have been obtained within any individual study.

DISCUSSION

The substantial proportion of nonsignificant cases strongly suggests that this database of Hawthorne studies is of generally poor quality. This may be a chance sampling, but may also be the result of a systematic tendency to mask poor quality research projects by the superficial addition of special control groups.

These results should not have been surprising. Cook's (1967) examination of Hawthorne studies anticipated many of these same results. He reported a table of designs somewhat similar to Table 1. Through a simpler meta-analytic voting procedure he placed available Hawthorne studies into one of three categories within each design: treatment greater than, equal to, or less than the control. Although he had very few studies with control group designs, only 4 with NTC and only 10 altogether, the treatment effect was nonexistent in the vast majority of the studies and greater than controls in only three studies!

The observation of some appropriate Hawthorne control cases, that is of the ideal and strong type, particularly within the unpooled effect sizes of individual observations, provides some direction for a more detailed examination of systematic differences in the design of these procedures. Unfortunately our preliminary examination of the distribution of these ideal cases was not encouraging: they were virtually equally distributed across the three Hawthorne procedures and there was no discernable pattern in their distribution across the types of Hawthorne activity.

Of course, there are further steps yet to be taken in our meta analysis of the current data. These analyses should lead us back to the original classification of variations of design practices and to examine their systematic variation in weighted effect sizes. Also, the present project was somewhat constrained by the requirement that the studies included in the database must have reference to Hawthorne. We have accumulated many more educational experiments that employ comparable design features yet make no reference to Hawthorne (Adair & Sharpe, 1986). Expansion of the project to this larger database, conceivably with controls for study quality, might provide a more adequate sampling of appropriate cases for analysis and interpretation.

Although our analysis is not complete, these preliminary results are not inconsistent with the earlier conclusion (Cook, 1967) of confusion surrounding the use of such artifact control procedures, and the lack of significant Hawthorne effects, and with my earlier view (Adair, 1984) that, because performance

expectancy or hypothesis awareness should be the artifact of primary concern, none of the control procedures may be appropriate to address Hawthorne concerns.

References

Adair, J. G. (1984). The Hawthorne effect: A reconsideration of the methodological artifact. Journal of Applied Psychology, 69, 334-345.

Adair, J. G., & Sharpe, D. (July, 1986). Placebo control practices in educational and psychotherapy research. Paper presented at the International Association of Applied Psychology, Jerusalem, Israel.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. American Educational Research Journal, 5, 437-474.

Cook, D. L. (1962). The Hawthorne effect in educational research. Phi Delta Kappan, 44, 116-122.

Cook, D. (1967). The impact of the Hawthorne effect in experimental designs in educational research. (Report No. 0726). Washington, D.C.: U.S. Office of Education.

Hedges, L. V., & Olkin, I. (1985). Statistical method for meta-analysis. Orlando, FA: Academic Press.

Smith, M. L., Glass, G., & Miller, T. I. (1980). The Benefits of Psychotherapy. London: Johns Hopkins University Press.

Table 1

Design Strategies To Control for Hawthorne Effects

Procedure	Frequency
I. Matching Hawthorne Group (HG) with experimental group (EG) on:	
A. A single variable - Although other dimensions may be incidental and or implied, it is the variable below that defines the HG.	
1. Activity, to equalize time, effort, interest, novelty	41
2. Awareness of experiment participation - includes telling both EG and HG, or only HG, that they are in an experiment	13
3. Attention, contact, or interaction	8
B. Multiple manipulation and matching - More than one variable is manipulated or introduced into the HG to equate it to the EG or to generate a Hawthorne effect:	
1. Activity and attention	10
2. Awareness and activity	3
3. Awareness and motivation - motivation refers to any special instruction or activity that is given to the HG to arouse their level of enthusiasm, expectation for success, or enhance their performance.	1
4. Awareness, motivation and attention	2
5. Awareness, activity and motivation	1
6. Attention, motivation, and activity	1
II. Special treatment administered to Hawthorne group.	
A. Special attention (activity of HG non-existent or incidental).	2

III. Create second control group (CG) so earlier group becomes HG.

A. No-treatment control (NTC) established in another location or independent of experiment so that it can be compared with EG and HG, without the awareness of experiment participation, presumably a condition that makes HG a "flawed" NTC. 2

B. After experiment a second CG established to compare as a no treatment control (NTC) with the EG and HG (i.e. the CG established during experiment and presumably ware of experiment). 1

Table 2

Percentages of studies employing or not employing no-treatment controls and combinations of control group activity and research design strategy

Activity	No-Treatment Controls: Absent N=31				No-Treatment Controls: Present N=54				Total N=85				
	Hawthorne Procedure[a]												
	Att	Act	Awa	Oth	Att	Act	Awa	Oth	Att	Act	Awa	Oth	Total
Irrelevant	5	6	1	0	7	7	1	0	12	13	2	0	27
Treatment Resemblance	1	8	2	2	2	17	2	1	3	25	4	3	35
Treatment Element	0	2	2	0	1	0	7	0	1	2	9	0	12
Other	1	1	0	0	4	0	3	2	5	1	3	2	11
Total	7	17	5	2	14	24	13	3	21	41	18	5	85

[a]Hawthorne Procedures: Att = match treatment group's attention, contact, and interaction or give only special attention (activity incidental); Act= Match treatment group's activity in time, effort, interest, novelty etc.; Awa = make subjects aware they are in an experiment; Oth = Other, including instructions to subjects, no treatment, etc. Studies employing combinations of the variables have been collapsed into Att, Awa or Oth.

Table 3

Summary Table of Homogeneity Tests of the Effect Sizes Weighted by Observations Within Each Study

OT	N	ID	Q_{TC}	Q_{HC}	df	OT	N	ID	Q_{TC}	Q_{HC}	df
1	2	14	0.036	.061	1	10	81	232	0.166	0.004	2
2	3	433	0.585	.531	1			233	3.562		1
3	4	550	0.461	.162	3			426	54.974*	12.421	19
4	4	428	1.500	.878	1			539	5.515	0.685	11
5	3	121	1.218	1.006	1			580	0.395	0.400	5
6	24	48	0.003		1			584	49.361*	37.378*	15
		58	2.607		2			593	29.138	14.594	23
		129	0.002	0.056	1	11	10	2	2.004	0.153	1
		428	0.634	1.108	1			33	0.203	0.005	1
		569	14.626	6.448	6			48	0.637	0.073	1
		577	4.275	15.914	5			128	20.474*	0.193	1
7	12	118	0.257	0.019	1	12	23	24	2.070	0.058	1
		119	0.299		4			25	0.773	0.576	1
		569	1.547	0.269	2			27	2.503	0.345	1
8	68	118	4.632	1.381	3			233	0.670		1
								424	23.170*	2.723	2
		250	8.964	83.319*	7			432	0.277		1
		539	4.860	2.249	11			435	0.994	0.806	3
		592	120.636*	8.277	7			575	0.007		1
		593	9.943	8.738	31	13	8	128	0.343	0.514	2
9	15	33	0.677	0.043	1			579	0.182	0.420	2
		233	1.918	1.818	1						
		577	8.429	8.241	9						

*p > .05

OT = Outcome type; N = total number of observations within each outcome type; ID = particular study on which multiple observations were obtained; TC = Treatment vs. No-treatment Control; HC = Hawthorne vs. Control.

Table 4

Summary ANOVA Tables for Homogeneity Tests of the Effect Sizes Weighted by Study Within Outcome Type

		Treatment vs. Control			Hawthorne vs. Control		
OT	df_b, df_w	Q_B	Q_W	F	Q_B	Q_W	F
1	-	-	0.036	-	-	0.061	-
2	1,1	0.745	0.585	1.273	-	0.531	-
3	-	-	0.461	-	-	0.162	-
4	1,2	2.381	1.50	1.587	.391	0.878	.445
5	1,1	1.408	1.218	1.156	.241	1.006	0.239
6	16,7	37.712	22.147	1.703	15.696	23.526	0.666
7	7,4	8.293	2.103	3.943	7.567	0.288	26.274**
8	63,4	61.162	149.035	0.410	48.981	103.464	0.411
9	11,3	0.624	11.024	0.057	8.744	10.102	0.865
10	76,8	174.859	143.111	1.222	170.379	65.482	2.602
11	4,5	40.610	23.318	1.741*	29.579	0.424	69.762**
12	12,10	345.391	11.024	31.331**	27.931	4.508	6.196*
13	4,3	20.564	0.525	39.169**	11.027	0.934	11.806*

OT=Outcome type; Q_B=Between-study chi-square under an OT; Q_W=Between-observation, within-study chi-square under an OT; - = Single study under an OT.

* $p < .05$

** $p < .01$

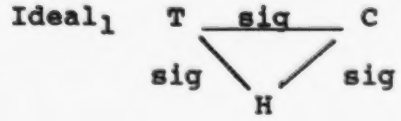
Table 5

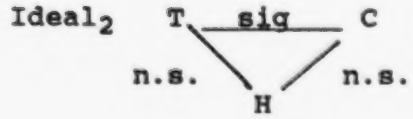
Frequency Counts of Significance Test Results for Each Case of Treatment-Hawthorne-Control Combination Across Homogeneous Outcome Types

	Outcome Type[b]														
Case[a]	1	2	3	4	5	6	7	8	9	10	11	12	13	Total	Percent
Ideal Case$_1$	0	0	0	0	0	0	0	2	0	1	0	4	0	7	3
Ideal Case$_2$	0	0	0	1	0	2	4	3	1	10	1	0	1	23	9
Strong Hawthorne	0	0	0	0	0	3	0	2	0	6	0	1	1	13	5
Weak Hawthorne	0	0	0	0	1	2	0	2	1	4	2	4	0	16	6
Rare Cases	0	0	0	0	0	2	0	3	3	0	3	0	1	12	5
Nonsignificant Cases	2	3	4	3	2	15	8	56	10	60	4	14	5	186	72

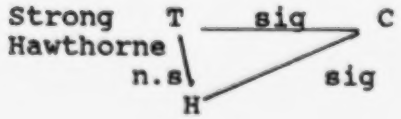
[a]See Figure 1 for the definition of each case.
[b]Outcome type: 1=fear/anxiety measures; 2=adjustment/school behavior scales; 3=sociopathic measures; 4=self-esteem scales; 5=personality scales; 6=behavioral measures; 7=intelligence tests; 8=academic aptitude tests; 9=self/peer/teacher ratings; 10=reading measures; 11=academic performance; 12=concept acquistion; 13=attitude assessment

Figure 1. Eight possible cases arising from significance test results on the unbiased effect sizes for treatment-Hawthorne-control comparisons.

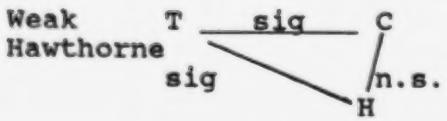




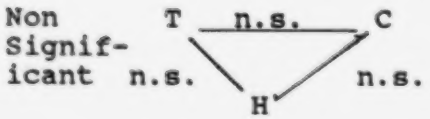
These are referred to as ideal cases in the sense that the Hawthorne group has been appropriately treated so that it results in performance scores midway between Treatment and NTC groups.



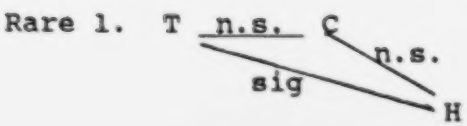
Hawthorne treatment effect was comparable to the Treatment effect



Hawthorne treatment was so weak that it was essentially identical to the NTC.



A totally non-significant study. The magnitude of the Treatment and Hawthorne treatment was too weak, or error was too great to show any significant effects.



Rare types result when there are nonsignifiant treatment effects, but unexpected effects associated with the Hawthorne group.